

# RESPECT: Reinforcement Learning based Edge Scheduling on Pipelined Coral Edge TPUs

Jiaqi Yin  
University of Utah  
Salt Lake City, US  
jiaqi.yin@utah.edu

Yingjie Li  
University of Utah  
Salt Lake City, US  
yingjie.li@utah.edu

Daniel Robinson  
University of Utah  
Salt Lake City, US  
u0714849@uemail.utah.edu

Cunxi Yu  
University of Utah  
Salt Lake City, US  
cunxi.yu@utah.edu

**Abstract**—Deep neural networks (DNNs) have substantial computational and memory requirements, and the compilation of its computational graphs has a great impact on the performance of resource-constrained (e.g., computation, I/O, and memory-bound) edge computing systems. While efficient execution of their computational graph requires an effective scheduling algorithm, generating the optimal scheduling solution is a challenging *NP-hard* problem. Furthermore, the complexity of scheduling DNN computational graphs will further increase on pipelined multi-core systems considering memory communication cost, as well as the increasing size of DNNs. Using the synthetic graph for the training dataset, this work presents a reinforcement learning (RL) based scheduling framework RESPECT, which learns the behaviors of optimal optimization algorithms and generates near-optimal scheduling results with short solving runtime overhead. Our framework has demonstrated up to  $\sim 2.5\times$  real-world on-chip inference runtime speedups over the commercial compiler with ten popular ImageNet models deployed on the physical Coral Edge TPUs system. Moreover, compared to the exact optimization methods, the proposed RL scheduling improves the scheduling optimization runtime by up to  $683\times$  speedups compared to the commercial compiler and matches the exact optimal solutions with up to  $930\times$  speedups. Finally, we perform a comprehensive generalizability test, which demonstrates RESPECT successfully imitates optimal solving behaviors from small synthetic graphs to large real-world DNNs computational graphs.

## I. INTRODUCTION

To efficiently utilize hardware platforms, scheduling algorithms implemented in deep learning compilers are critical in deploying these hyper-dimensional computationally-intensive workloads, and this is a classical *NP-hard* combinatorial optimization (CO) problem. Mostly, vendor-specific libraries rely on hand-crafted domain-specific heuristics to optimize the executions, which trades the execution performance for scheduling solving time. Specifically, the limitations of the existing scheduling approach can be summarized as follows: (1) existing algorithms are either heuristics that lack in quality of optimization or exact/brute-force algorithms lacking in scalability [5], [12], [21], [24]. The challenges for large-scale deep learning executions are still rising, particularly for edge devices; (2) hand-crafted heuristics can be efficient but the development process requires high engineering efforts and domain knowledge for compilation and hardware systems; (3) there have recently seen machine learning (ML) based frameworks for design automation [13], [22], [23] and compilation [4], [9], [14]. However, they are either limited to certain graph structures/sizes or require online training or retraining [4], [9], [14]. More importantly, existing ML-based approaches all rely on real-world data collection, which is time-consuming and limited in generalizability [11], [13], [15].

This work presents RESPECT, a novel combinatorial reinforcement learning framework that imitates the behaviors of graph combinatorial algorithms for scheduling DNN computational graphs on edge devices. RESPECT aims to perform near-optimum scheduling on edge computing platforms with significantly reduced runtime complexity compared to the exact or brute-force scheduling algorithms. Moreover, the proposed RL model is generalizable to arbitrary DNN computational

graphs within the same edge device domain without retraining. The contributions of this work can be summarized as follows: (1) to achieve the near-optimal scheduling performance with better scalability, we propose a novel RL-based scheduling framework that imitates the algorithmic behaviors of existing optimal scheduling algorithm, with data-independent training setups, i.e., training conducted fully on synthetic datasets; (2) the experimental results are conducted on a physical multi-stage pipelined Edge TPU system [3], [20] via USB 3.0 communication interface which demonstrates significant Edge TPU inference runtime speedups over the commercial Edge TPU compiler with ten popular ImageNet DNNs models. Simultaneously, RESPECT offers up to  $683\times$  speedups in scheduling solving time, and up to  $2.5\times$  performance speedup; and (3) finally, we compare the proposed RL method to the exact optimal scheduling method, where the proposed RL approach offers up to  $930\times$  speedups for scheduling, with a near-optimal Edge TPU inference runtime performance. Our results and Edge TPU deployment infrastructure are released publicly [1].

## II. BACKGROUND

In the modern deep learning frameworks, machine learning algorithms are represented as computational graphs, where each graph is a directed graph  $G(V, E)$  with nodes  $V$  describing operations, and edges  $E$  representing the dataflows that connect the operators and input/output tensors. Practically, when deploying DNNs on hardware accelerators, the computational graphs are mostly represented as directed acyclic graphs (DAG), where the acyclic paths are unrolled to maximize the hardware performance. The computational graphs are mostly generated with static compilation.

Specifically, the optimization objectives of scheduling computational graphs can be defined as follows: **Given:** (1) A DAG  $G(V, E)$  where  $V$  represents the set of operations in the DNN computational graphs, and  $E$  represents the set of edges; (2) A set of scheduling constraints, which may include dependency constraints, resource constraints, execution time, memory allocations, etc. **Objective:** Construct an exact optimal schedule  $S = s_0, s_1, \dots, s_n$ ,  $n \leq |V|$ , where  $n$  represents the number of scheduling stages. The operation set in the computational graph  $V$  will be allocated to  $S$  that satisfies all scheduling constraints. For example, in a multi-stage pipelined Edge TPU system in Figure 1a, the resulted schedule assigns computation node  $N_i$  to  $s_0$  (Edge TPU:0),  $N_k$  to  $s_1$  (Edge TPU:1),  $N_l$  and  $N_o$  to  $s_2$  (Edge TPU:2),  $N_j$ ,  $N_p$  and  $N_q$  to  $s_3$  (Edge TPU:3), etc.

Inference performance on edge devices is highly sensitive to scheduling solutions of computational graphs, due to the limited compute elements, memory, and communication bandwidth. The optimization problem of computational graph scheduling on edge devices is similar to the classic resource-constrained scheduling (RCS) problem, which has been widely studied. These studies have resulted in a set of

<sup>1</sup><https://github.com/Yu-Utah/RESPECT>

heuristic scheduling methods (Hu’s Algorithm), List Scheduling, and Force-Directed Scheduling [12], [19]. It is believed that Google Edge TPU compiler adopts the heuristic methods to schedule computation graphs with an acceptable schedule solving time. Although heuristic methods can easily obtain the solution, the scheduling solutions could be far from optimum. Another viable scheduling method is iterative metaheuristics, such as simulated annealing, ant colony, and dynamic programming optimizations [10]. For example, [1] proposed dynamic programming based adaptive budgeting scheduling technique, which can optimize scheduling efficiently without either performance guarantees or significant long optimization runtime requirements.

On the other hand, the RCS problem can be solved with exact optimal methods using modern formal method solvers such as SAT and SMT solvers, or constraint solving such as Integer Linear Programming (ILP), after it maps to a constraint satisfaction problem which consists of logical connectives of linear constraints [8], [21], [24]. It integrates specialized solvers with propositional satisfiability search techniques to achieve conflict-driven learning [6]. Even though, such exact methods can produce the optimal solution, it has a drawback of high schedule solving time, making it difficult to apply to super large-scale computation graph scheduling. Thus, existing RCS scheduling methods have a clear trade-off between scheduling solving time complexity and the scheduling solution quality. This work aims to push the Pareto-frontier of traditional scheduling methods specifically for ML inference on edge devices, to offer near-optimum scheduling solutions at the optimization cost of fast heuristics.

### III. APPROACH

Figure 1a shows the overview of RESPECT framework – Step ①: RESPECT takes the real-world DNN model as input and converts it to a hypergraph depending on the computation nodes and dependency between nodes. Step ②: The embedding of the computation graph is generated in RESPECT, which includes node attributes (topological level, node’s ID), node dependency (parents topo level, parents’ IDs), and memory consumption (node’s memory). Step ③: A LSTM-PtrNet trained with reinforcement learning (RL) is employed to generate the schedule sequence with the embedded input queue  $q$  from step ②. Note that the LSTM-PtrNet model is trained with synthetic datasets to save the efforts for real-world data collection. Step ④: The model is ready to be deployed to devices with the scheduling sequence from step ③. Additionally, post-inference processing with domain-specific constraints is applied to make the deployment compatible with physical devices. For example, RESPECT will quantize the Tensorflow Model and deploy sub-models to real-world hardware devices using the Tensorflow Toco converter.

#### A. DNN computational Graph Embedding

For learning purposes, the computational graph is usually embedded into vector space. Considering the performance of scheduling on edge devices, the embedding of computational graph consists of four components as shown in Figure 1a. (1) *absolute coordinates* generated based on topological ordering. For example, for node  $N_i$ , its absolute coordinates will be  $T_i$  in the first column in its embedding. Note that there are many variants of topological orders of a given DAG. In this work, we use As-Soon-As-Possible (ASAP) ordering, where each node is ordered as close to the source node as possible; (2) *relative coordinates* consisting of parent nodes absolute coordinates and parents IDs. For example, for node  $N_j$ , the relative coordinates will be its parents topological level  $T_i$  and the parents IDs  $N_i$ . For source nodes, absolute coordinates are set to 0, and their IDs are set to  $-1$ ; (3) node IDs are unique integers generated by hashing all the operator names. For node  $N_i$ , its node’s ID will be  $N_i$ ;

(4) cache and memory locality are one of the most critical metrics in DNNs acceleration. Thus, we add the memory consumption of each operator in the last embedding column. Specifically, absolute coordinates embedding represents the topological order of every node which mimics the feasible scheduling space for a given node. Relative positions are encoded using both the absolute coordinates and node IDs. Intuitively, relative coordinates maintain the dependency constraints and absolute coordinates. For example, in Figure 1a, the data dependency constraints for  $N_p$  are the complete scheduling of its parent nodes, i.e.,  $N_i$  and  $N_o$ . Meanwhile,  $N_j$  and  $N_p$  can be scheduled together on available computation resources as there is no data dependency between them. Besides, the absolute coordinates of node  $N_j$  and its parent node  $N_i$  are 2 and 1, respectively. Thus, the valid schedule stages for  $N_j$  can be calculated, i.e., any available stage between the absolute coordinates of  $N_i$  and  $N_j$ . When  $N_i$  is deployed to  $s_0$ , available stages for deploying  $N_j$  include  $\{s_0, s_1, s_2, s_3\}$ .

#### B. Formulations and Neural Architecture

As discussed earlier, brute-force and exact methods scheduling algorithms can generate the best scheduling solution search [5], [8], [24] at the cost of high runtime complexity, which fails in scaling up to large problems. While heuristic algorithms optimize the schedules more efficiently at scale, they suffer from the quality of results. Thus, we aim to develop an RL framework that imitates the algorithmic behaviors of any optimal scheduling algorithm (e.g., exact or brute-force), such that it performs polynomial time scheduling with near-optimum results at inference runtime.

**RL Formulation for DNN computational Graph Scheduling** – Given a computational graph DAG  $G(V, E)$ , the RL-agent is trained to develop a policy  $\pi$ , which picks computation nodes in the same order as a target exact or brute-force algorithm does. In this work, we define a sequence order generated by our method  $\pi(i), i \leq |V|$ , and a sequence order figured by a given deterministic exact scheduling method as  $\gamma(i), i \leq |V|$ . This exact scheduling method can be formulated using ILP, which offers the exact optimal schedule w.r.t a given optimization objective. The reward function is designed as the similarity comparison between the two sequences:

$$R = \frac{\sum(\pi(i) \cdot \gamma(i))}{\max(\sqrt{\sum \pi(i)^2} \cdot \sqrt{\sum \gamma(i)^2}, \epsilon)} \quad (1)$$

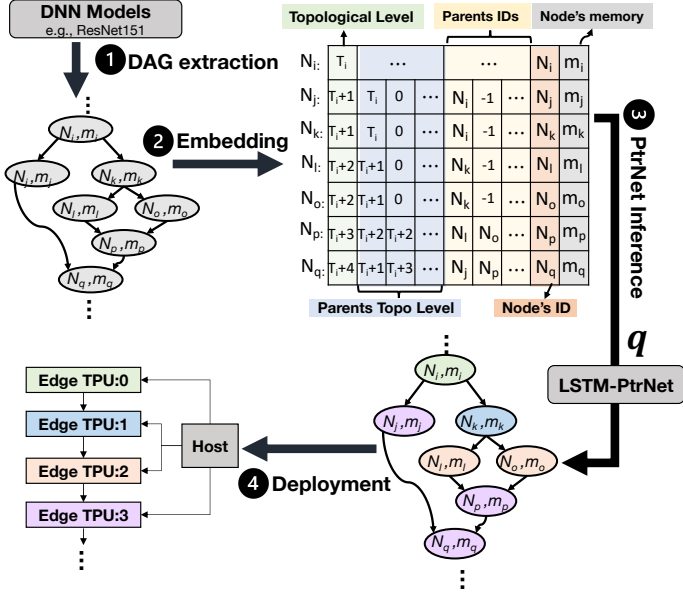
where  $\epsilon$  is a small constant number to guarantee the validity of the equation. Let  $N_i \rightarrow s_k, i \in [1, |V|], k \in [0, n]$  be the solution of scheduling node  $N_i$  at  $s_k$ . Let  $S' = \{s_0, s_1, \dots, s_n\}$ , where  $n$  is the number of Edge TPU stages, be the sequence produced by our policy  $\pi(i)$ , i.e., the output of the RL agent for a given computational graph. The targeted scheduling solution generated by the exact scheduling method is the ground truth label sequence, denoted as  $S$ .  $S'$  and  $S$  can be computed as follows:

$$S' = \rho(\pi(i), s_k); \quad S = \rho(\gamma(i), s_k), i \in [1, |V|], k \in [0, n] \quad (2)$$

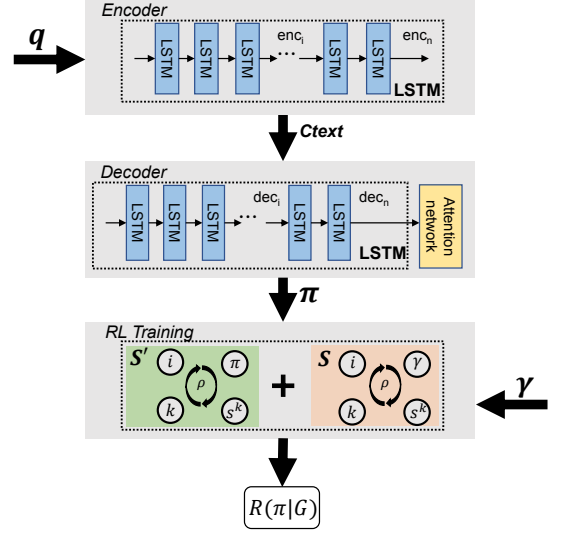
where  $\rho$  denotes the scheduling algorithm w.r.t the specific Edge TPU. Hence, to imitate the behaviors of the targeted exact scheduling method, the reward is designed using cosine similarity between the generated schedule sequence  $S'$  and the ground truth sequence  $S$  (Equation 3).

$$R = \frac{\sum(S_i^k \cdot S'^k_i)}{\max(\sqrt{\sum S_i^{k^2}} \cdot \sqrt{\sum S'^k_i{}^2}, \epsilon)}, i \in [1, |V|], k \in [0, n] \quad (3)$$

While maximizing the reward function  $R$ , parameters of the stochastic policy  $p(\pi|G)$  will be optimized to assign high probabilities to



(a) RESPECT framework overview – Step 1: real-world DNN model DAG extraction; Step 2: graph embedding based on topological level, node's ID and memory consumption; Step 3: inference through the encoder-decoder based LSTM-PtrNet graph scheduler; Step 4: hardware deployment.



(b) Detailed architecture of the LSTM-PtrNet model. The encoder network will digest the embedded input queue  $q$  and transform it into a sequence of context  $C_{text}$ , and produce a sequence of latent memory states  $enc_i$ . The decoder network will produce the predicted sequence  $\pi$ , and generate the latent memory state  $dec_i$ . RL model training is based on sequence  $\pi$  and ground truth label  $\gamma$ .

Fig. 1: Overview of our RL-based scheduling framework.

#### Algorithm 1: Pointing Mechanism Decoding Flow

**Initialization:**  $d \leftarrow$  decoder input;  $C = \{C_{text}_i\}_{i=1}^{|V|}$ ;  $dec_0$ ;  $\{emb_i\}_{i=1}^{|V|}$ ;  $\theta, \omega, \beta$  (trainable parameters);  
**for**  $i = 1$  **to**  $n$  **do**  
     $h, dec_i \leftarrow \text{Dec}(d, dec_{i-1})$ ;  
     $h \leftarrow \text{glimpse}(C \cdot \theta_g, \omega_g \cdot h + \beta_g)$ ;  
     $P^i \leftarrow \text{pointer}(\tanh(C \cdot \theta_p, \omega_p \cdot h + \beta_p))$ ;  
     $idx_i \leftarrow \text{argmax}_{j=1}^n \frac{\exp P_j^i}{\sum_{j=1}^n \exp P_j^i}$ ;  
     $d \leftarrow emb_{idx_i}$   
**end**  
 $S' \leftarrow \rho(\pi(i), s_k)$ ;  $S \leftarrow \rho(\gamma(i), s_k)$ ,  $i \in [1, |V|]$ ,  $k \in [0, n]$ ;

sequence order closer to the target sequence. The chain rule utilized to factorize the sequence probability distribution can be expressed as:

$$p(\pi|G) = \prod_{i=1}^{|V|} p(\pi(i)|\pi(<i), G) \quad (4)$$

where  $\pi(<i)$  represents the output sequence generated by previous epochs. At each epoch,  $p(\pi|G)$  is updated based on input computation graph and the previous sequence  $\pi(<i)$ .

**RL Agent Architecture** – We extend pointer network (PtrNet) architecture [2] as our RL agent. PtrNet excels in finding the path with target objectives to be optimized. It achieves huge success in solving some combinatorial problems over graphs, such as the Traveling Salesman Problem [2]. With the advantage of attention mechanism [16], PtrNet reinforces the dependency constraints among nodes and overcomes the limitations of learning combinatorial graph algorithms with fixed size of graph inputs. Note that in order to fully evaluate the RL scheduling performance on physical edge devices, the results need to satisfy domain-specific hardware execution requirements, such as data dependency.

The RL agent architecture consists of encoding and decoding components, as shown in Figure 1b. Specifically, each component is a Long Short-Term Memory (LSTM) cell. **Encoder network** – The encoder network digests the encoding input queue  $q$  of nodes and transforms it into a sequence of context  $\{C_{text}_i\}_{i=1}^{|V|}$  representing the information of each node, where  $C_{text}_i \in R^d$ ,  $d$  is the hidden dimension of LSTMs. The concatenation of all contexts will be used as reference matrix  $C$  in picked nodes during decoding step,  $C(i) = C_{text}_i$ ,  $i \in [1, |V|]$ . When using LSTM as an encoder, it produces a sequence of latent memory states  $\{enc_i\}_{i=1}^{|V|}$  recording the propagation encoding message from the first node to current one along with the contexts, where the final state  $enc_{|V|}$  is the initial latent memory state for decoding. **Decoder network** – When using LSTM as decoder network, it generates its own latent memory state  $dec_i \in R^d$  at each decoding step  $i$  to update the previous one. We illustrate the detailed decoding procedure in Algorithm 1. With context reference matrix  $C$  from encoder, decoder will produce a selection probability distribution over candidate computation nodes using pointing mechanism. The LSTM is employed as the decoder (Dec), which takes the node embedding  $d$  and the last latent memory state  $dec_{i-1}$  as inputs to produce the annotation  $h$ , and updates the latent memory state  $dec_i$ . Then attention networks including *glimpse* and *Ptr-Net* [17] are employed to produce the sequence order  $\pi$ . Specifically, *glimpse* updates  $h$  with the context matrix  $C$ , and *Ptr-Net* will produce the probability over candidates with the hidden state information from both encoder ( $enc_i$ ) and decoder ( $dec_i$ ). The next node will be selected with softmax. Once a computation node  $N_{idx_i}$  with the highest probability is picked, its embedding  $emb_{idx_i}$  will be passed as the input  $d$  to LSTMs together with the information of the last latent memory state  $dec_{i-1}$  during the next decoding step. The logits of the nodes that appeared in the solution  $\pi$  are set to  $-\infty$  in the network to guarantee the validity of the solution. The input to the first decoding step  $dec_0$  is a trainable parameter sharing the same



dimension with node embedding.

**RL Training** – We use the model-free policy-based RL training method to optimize the parameters of a pointer network denoted among  $\theta$ . The learning objective is the expected similarity of node distribution. Given an input graph  $G$ , the optimization objective is to maximize the following cosine similarity reward:

$$J(\theta|G) = \mathbb{E}_{\pi \sim p_{\theta}(\cdot|G)}(1 - R(\pi|G)) \quad (5)$$

In this work, we deploy policy gradient algorithms and stochastic gradient descent algorithms to optimize the parameters [18]. Reward as a score to evaluate the resulted distribution probability is applied as a coefficient in gradient calculation, such that the gradient of Equation 5 is constructed as follows:

$$\nabla_{\theta} J(\theta|G) = \mathbb{E}_{\pi \sim p_{\theta}(\cdot|G)}[(1 - R(\pi|G) - b(G)) \nabla_{\theta} \log p_{\theta}(\pi|G)] \quad (6)$$

where  $b(G)$  represents a baseline in LSTM-PtrNet model that performs the best among all the past training iterations, i.e., rollout baseline [7], and it is applied in loss function to reduce the variance of the gradients. Note that, unlike existing RL-based scheduling works, our framework imitates the behaviors of any scheduling algorithm, such that the resulted RL agent can perform as generally as the imitated deterministic algorithm.

**Synthetic training dataset** – The limited number of computation graphs for training the RL models restricts the generalizability of RL in solving graph-based combinatorial optimization problems. Besides, to simultaneously minimize the training efforts, the training dataset needs to balance data coverage and graph size (i.e.,  $|V|$  and  $|E|$  in  $G(V, E)$ ). There is a critical advantage to use synthetic data sampling during RL training: synthetic graph sampler has full control over graph complexity, and memory attributes of operators, which offers much better data coverage over real-world DNN computation graphs. Therefore, we use the synthetic graphs only in training, where we integrate a DAG sampler into our RL training framework which randomly generates network graphs with  $|V| = 30$  but with different graph complexities. The synthetic datasets are designed to mimic the structures and properties of DNN computational graphs. The complexities of the trained graphs can vary with different graph degrees ( $\deg(V) \in \{2, 3, 4, 5, 6\}$ ), which is the maximum number of incoming edges connecting to the vertices  $V$  in graph  $G(V, E)$ . RESPECT is trained with 1 million random graphs, with 200,000 graphs for each  $\deg(V)$  in  $\{2, 3, 4, 5, 6\}$ .

**Post-Inference Processing** – Unlike TSP algorithms, DNN computational graph scheduling needs to satisfy domain-specific constraints to successfully deploy the scheduled graphs on hardware platforms. Thus, a post-inference processing procedure is added to our RL framework, which is executed at the deployment stage, and takes the inference output  $S'$  as input. To be specific for our evaluation on Edge TPU platforms, this procedure corrects the dependency violation by simply pushing the involved node forward, which is a deterministic step with minimum changes to the RL solution. Besides, Edge TPU hardware requires children nodes of any node to be in the same pipeline, where the post-inference procedure assigns these nodes to the earliest predicted stage.

#### IV. EXPERIMENTAL RESULTS

In this section, we aim to experimentally demonstrate the effectiveness of the proposed RL-based scheduling approach which offers near-optimal scheduling solutions with runtime cost of inference, using real-world physical Edge TPU devices. The two comparison baselines include commercial Edge TPU compilers (e.g., Google’s Edge TPU

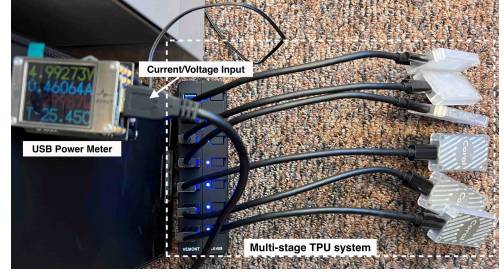


Fig. 2: Illustration of our multi-stage pipeline Edge TPUs and energy efficiency evaluation system via USB 3.0 communication interface.

compiler<sup>2</sup>), which adopts the heuristic method to schedule computation graphs, and an exact-optimal scheduling method conducted on constraint solving scheduling using ILP solver [8], [24]. We evaluate RESPECT in three aspects: 1) on-chip inference runtime, 2) solving runtime, and 3) the optimization gaps to exact optimal solutions.

Our framework integrates RESPECT scheduling with Tensorflow-Lite (TFLite) and Edge TPU Compiler to complete the deployment flow. It takes single or multiple DNN models and the number of pipeline stages as inputs, and outputs  $n$  partitioned subgraphs for deployment on Edge TPU devices. The framework involves graph construction, solving via RESPECT, solution extraction, and deployment.

**Experiment and evaluation setups** – We use ten popular ImageNet classification models (Table I) as the benchmarks. Training, inference, and explorations of the proposed RL methods are conducted on one Nvidia 2080 Ti with Intel Xeon Gold 6230 x20 CPUs. For the exact method, the optimization ILP formulations are solved using IBM ILOG CPLEX. The scheduling solution with heuristic method is acquired from Google’s Edge TPU compiler<sup>1</sup>. RESPECT is implemented with PyTorch. The neural architecture encoder and decoder are built with LSTMs with 256 cells. Training is conducted on 300 epochs with the learning rate of  $10^{-4}$  and batch size of 128 using Adam optimizer. Besides, we build a central-hosted EdgeTPU system for experiment results evaluation which is shown in Figure 2.

TABLE I: Statistics of DNNs models and their computational graphs used for evaluating inference runtime on Pipelined Edge TPU Systems.

|           | Xception    | ResNet50          | ResNet101   | ResNet152   |
|-----------|-------------|-------------------|-------------|-------------|
| $ V $     | 134         | 177               | 347         | 517         |
| $\deg(V)$ | 2           | 2                 | 2           | 2           |
| Depth     | 125         | 168               | 338         | 508         |
|           | DenseNet121 | ResNet101v2       | ResNet152v2 | DenseNet169 |
| $ V $     | 429         | 379               | 566         | 597         |
| $\deg(V)$ | 2           | 2                 | 2           | 2           |
| Depth     | 428         | 371               | 558         | 596         |
|           | DenseNet201 | InceptionResNetv2 |             |             |
| $ V $     | 709         | 782               |             |             |
| $\deg(V)$ | 2           | 4                 |             |             |
| Depth     | 708         | 571               |             |             |

##### A. Edge TPU Inference Runtime Evaluation

We compare the scheduling results of RESPECT with the heuristic method based Edge TPU compiler and exact methods based ILP solver. The results included in Figure 4 are the average runtime of 10 rounds of 1,000 ImageNet inferences. The vertical axis represents the inference runtime results normalized w.r.t the runtime obtained using Edge TPU compiler.

As shown in Figure 4 the proposed RL scheduler consistently outperforms the commercial Edge TPU compiler. Specifically, the RL scheduler provides  $1.06\times$ ,  $1.08\times$ ,  $1.65\times$  speedups over Edge

<sup>2</sup><https://coral.ai/docs/edgetpu/compiler/>

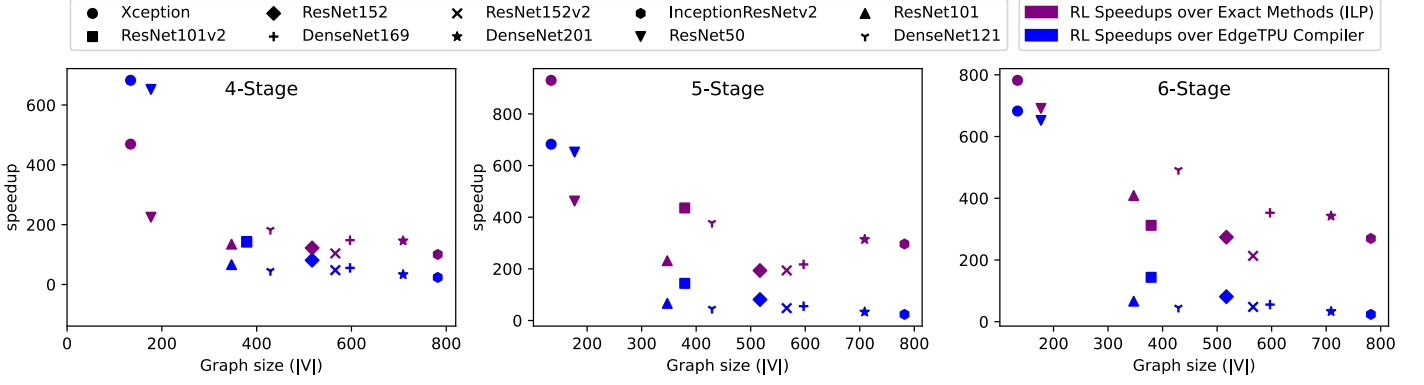


Fig. 3: Schedule solving time across ten real-world ImageNet models with different graph sizes. The horizontal-axis and vertical-axis represent graph sizes  $|V|$ , and speedups over the baselines, respectively. RL method offers  $24 - 683\times$  speedups over Edge TPU compiler and  $100 - 930\times$  over exact methods.

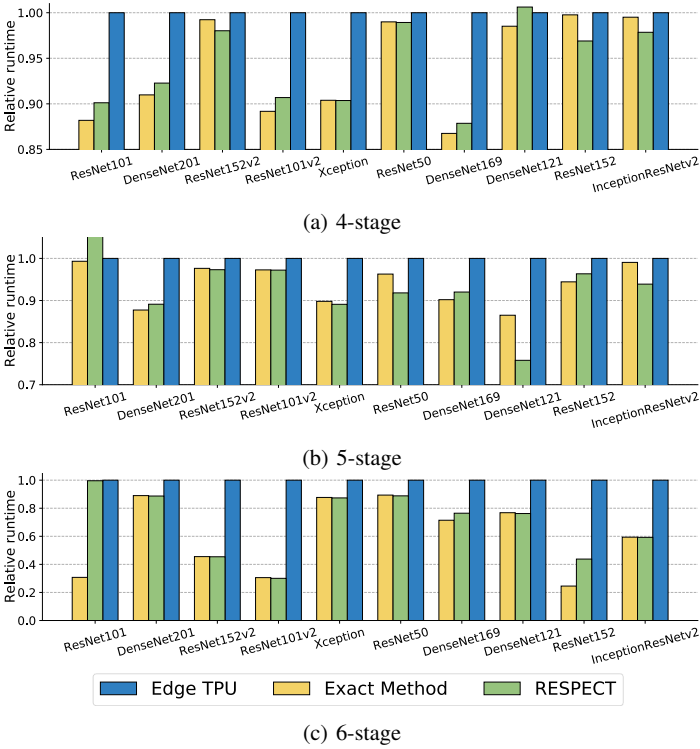


Fig. 4: Multi-stage pipelined Edge TPUs inference runtime comparisons between the proposed RL methods, exact methods, and commercial Edge TPU compiler (baseline scale=1). RL methods provide a near-optimum scheduling assignment. The runtime performance has been consistently improved over commercial compilers with 4-, 5-, and 6-stage pipelined Edge TPU systems (e.g., ResNet101v2 and ResNet152 execute  $2.5\times$  faster than the Edge TPU compiler).

TPU compiler in 4-, 5-, and 6-stage pipelining options. For example, ResNet101v2 and ResNet152 execute  $2.5\times$  faster than Edge TPU compiler in 6-stage pipelining. Second, compared to the scheduling solutions generated with the exact optimal scheduling method solved by ILP, the proposed RL approach demonstrates near-optimum Edge TPU inference performance, but with a significantly reduced runtime in scheduling. Detailed discussions on solving time complexity are shown in Section IV-B. Third, the average Edge TPU inference runtime

improvements of the proposed RL approach over Edge TPU compiler increase as the number of pipeline stages increases. For example, our proposed RL method provides a speedup of  $1.03\times$  for ResNet152 in a 4-stage system, and the inference runtime speedup increases to  $2.28\times$  in a 6-stage system. The main reason is that as the number of stages increases, the complexity of the scheduling optimization increases, where the performance of heuristics methods will further degrade.

**Performance modeling miscorrelation** – From the analysis of runtime performance shown in Figures 4, we observe that exact methods sometimes perform worse than RESPECT. This is due to the miscorrelation between the high-level abstract modeling of computational graphs and the actual systolic hardware specifications, in which the hardware architecture and backend compilation are closed source. Such miscorrelations of the high-level performance modeling can be used to guide the scheduling optimizations but introduce optimization noise to the actual on-chip performance. Specifically, in this work, both the exact method and RESPECT optimize the DNN model scheduling from the aspects of the memory allocation and communication cost. Note that directly modeling or estimating the runtime from the computation graph and its schedule is impossible given the complexity of the Edge TPU compute components and caching/memory system [3].

## B. Solving Scalability Evaluation

In this section, we further compare the schedule solving runtime of our proposed RL framework with the same set of baselines on the same set of DNN models. The results are shown in Figure 3 where the horizontal-axis represents the graph sizes, i.e.,  $|V|$  in the computation graph  $G(V, E)$ . The vertical-axis shows the speedup in solving runtime of the RL framework over the two baselines.

Three important conclusions can be summarized from Figure 3: (1) **Compared to the heuristic methods (Edge TPU compiler), the main advantage of the proposed RL approach is the improved quality of scheduling in Edge TPU inference runtime.** Specifically, our RL approach offers  $24\times$  to  $683\times$  speedups over Edge TPU compiler, in generating the scheduling solutions, with up to  $2.5\times$  inference runtime improvements discussed in Section IV-A. For example, for ResNet152, the RL generated schedule offers  $2.4\times$  speedups in Edge TPU inference runtime and simultaneously improves the scheduling solving time by  $85\times$ . (2) **Compared to the exact method conducted on constraint solving scheduling, our RL approach offers  $100\times$  to  $930\times$  speedups in generating scheduling solutions, where the Edge TPU inference runtime is almost identical.** For example, for

the Xception model, the Edge TPU inference runtime is almost the same for RL approach and the exact method, while the RL scheduling solving time is  $930\times$  faster than the exact method.

### C. Gap-to-Optimality Analysis

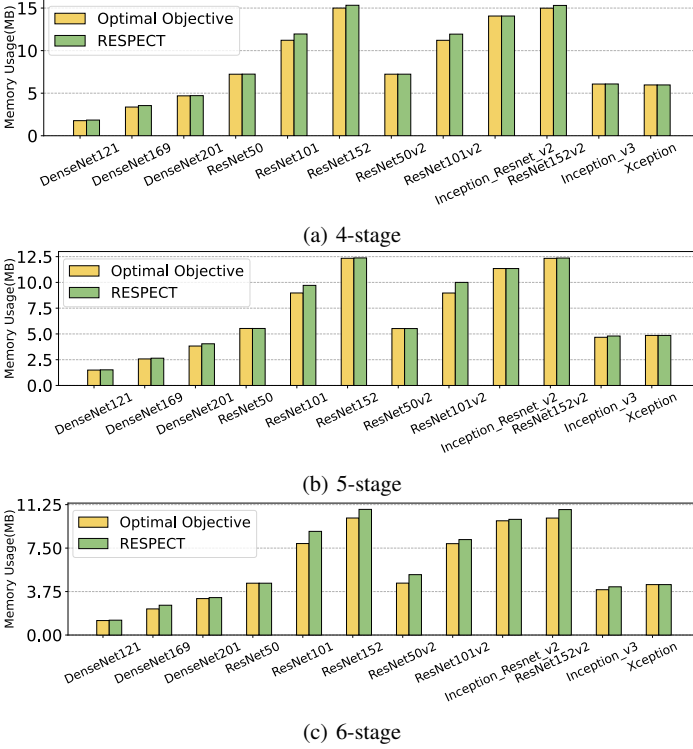


Fig. 5: Gap-to-optimal analysis between RESPECT and exact scheduling. (a)–(c): Parameter caching results of ImageNet DNNs models cross 4-, 5-, 6-stage pipeline settings.

Finally, we perform a generalizability analysis to demonstrate the capabilities of generating near-optimal scheduling solutions of the RESPECT framework. We confirm the generalizability of our RL framework by evaluating the following metrics: quality-of-results w.r.t the exact optimal solutions, namely *gap-to-optimal* of RESPECT inference. Specifically, we collect the parameter caching values, including on-cache and off-cache memory caching values across all pipeline stages, and calculate the absolute differences in peak memory usage per stage between exact-optimal and RESPECT scheduling solutions. The results are conducted on the ImageNet DNNs models, shown in Figures 5a to 5c. We can see that RESPECT provides near-optimal results in parameter caching in comparison to exact optimal solutions, with only 2.26%, 2.74%, and 6.31% average performance gap-to-optimal, across all 4-, 5-, and 6-stage pipeline settings.

### V. CONCLUSION

This work presents a reinforcement learning based scheduling framework, which imitates the behaviors of optimal optimization algorithms in significantly reduced runtime complexity. Our framework has demonstrated up to  $2.5\times$  runtime speedups over the commercial Edge TPU compiler, using ten popular ImageNet models on three different physical pipelined Google Edge TPU systems. More importantly, compared to the exact optimization methods solved by heuristics and brute-force, the proposed RL scheduling improves the scheduling solving time by several orders of magnitude.

**Acknowledgment** – This work is supported by National Science Foundation (NSF) under NSF-2047176, NSF-2019336, NSF-2008144, and NSF-2229562 awards.

### REFERENCES

- [1] B. H. Ahn, J. Lee, J. M. Lin, H.-P. Cheng, J. Hou, and H. Esmailzadeh, “Ordering chaos: Memory-aware scheduling of irregularly wired neural networks for edge devices,” *Proceedings of Machine Learning and Systems*, vol. 2, pp. 44–57, 2020.
- [2] I. Bello, H. Pham, Q. V. Le, M. Norouzi, and S. Bengio, “Neural combinatorial optimization with reinforcement learning,” *arXiv preprint arXiv:1611.09940*, 2016.
- [3] A. Boroumand, S. Ghose, B. Akin, R. Narayanaswami, G. F. Oliveira, X. Ma, E. Shiu, and O. Mutlu, “Mitigating edge machine learning inference bottlenecks: An empirical study on accelerating google edge models,” *arXiv preprint arXiv:2103.00768*, 2021.
- [4] H. Chen and M. Shen, “A deep-reinforcement-learning-based scheduler for fpga hls,” in *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2019, pp. 1–8.
- [5] J. Cong and Z. Zhang, “An Efficient and Versatile Scheduling Algorithm Based on SDC Formulation,” *Design Automation Conf. (DAC)*, 2006.
- [6] L. De Moura and N. Björner, “Satisfiability Modulo Theories: Introduction and Applications,” *Communications of the ACM*, 2011.
- [7] W. Kool, H. Van Hoof, and M. Welling, “Attention, learn to solve routing problems!” *arXiv preprint arXiv:1803.08475*, 2018.
- [8] C. E. Leiserson and J. B. Saxe, “Retiming synchronous circuitry,” *Algorithmica*, vol. 6, no. 1, pp. 5–35, 1991.
- [9] H. Mao, M. Schwarzkopf, S. B. Venkatakrishnan, Z. Meng, and M. Alizadeh, “Learning scheduling algorithms for data processing clusters,” in *Proceedings of the ACM Special Interest Group on Data Communication*, 2019, pp. 270–288.
- [10] G. D. Micheli, *Synthesis and Optimization of Digital Circuits*. McGraw-Hill Higher Education, 1994.
- [11] D. Pal, C. Deng, E. Ustun, C. Yu, and Z. Zhang, “Machine learning for agile fpga design,” *Machine Learning Applications in Electronic Design Automation*, pp. 471–504, 2022.
- [12] P. G. Paulin and J. P. Knight, “Force-directed scheduling for the behavioral synthesis of asics,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 8, no. 6, pp. 661–679, 1989.
- [13] H. Ren and J. Hu, *Machine Learning Applications in Electronic Design Automation*. Springer Nature, 2023.
- [14] S. Sheng, P. Chen, Z. Chen, L. Wu, and Y. Yao, “Deep reinforcement learning-based task scheduling in iot edge computing,” *Sensors*, vol. 21, no. 5, p. 1666, 2021.
- [15] E. Ustun, S. Xiang, J. Gui, C. Yu, and Z. Zhang, “Lamda: Learning-assisted multi-stage autotuning for fpga design closure,” in *2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2019, pp. 74–77.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [17] O. Vinyals, M. Fortunato, and N. Jaitly, “Pointer networks,” in *Advances in neural information processing systems*, 2015, pp. 2692–2700.
- [18] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.
- [19] T. Yang and A. Gerasoulis, “List scheduling with and without communication delays,” *Parallel Computing*, 1993.
- [20] A. Yazdanbakhsh, K. Seshadri, B. Akin, J. Laudon, and R. Narayanaswami, “An evaluation of edge tpu accelerators for convolutional neural networks,” *arXiv preprint arXiv:2102.10423*, 2021.
- [21] J. Yin, Z. Zhang, and C. Yu, “Exact memory-and communication-aware scheduling of dnn on pipelined edge tpus,” in *2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)*. IEEE, 2022, pp. 203–215.
- [22] C. Yu, H. Xiao, and G. De Micheli, “Developing synthesis flows without human knowledge,” in *Proceedings of the 55th Annual Design Automation Conference*, 2018, pp. 1–6.
- [23] C. Yu and Z. Zhang, “Painting on placement: Forecasting routing congestion using conditional generative adversarial nets,” in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.
- [24] Z. Zhang and B. Liu, “SDC-Based Modulo Scheduling for Pipeline Synthesis,” *Int’l Conf. on Computer-Aided Design (ICCAD)*, 2013.